

The Politics of Vocabulary Control

Musings on schema.org and
Linked Open Data

Prof. Dr. Stefan Gradmann
Director University Library / Professor (Arts)
Stefan.Gradmann@kuleuven.be

The Menu

- Linked Open Data: a Cheerful Mess
- Schema.org: Order and Discipline
- The Hidden Agenda
- Privatised Censorship?

The Web: a Cheerful, Robust Mess

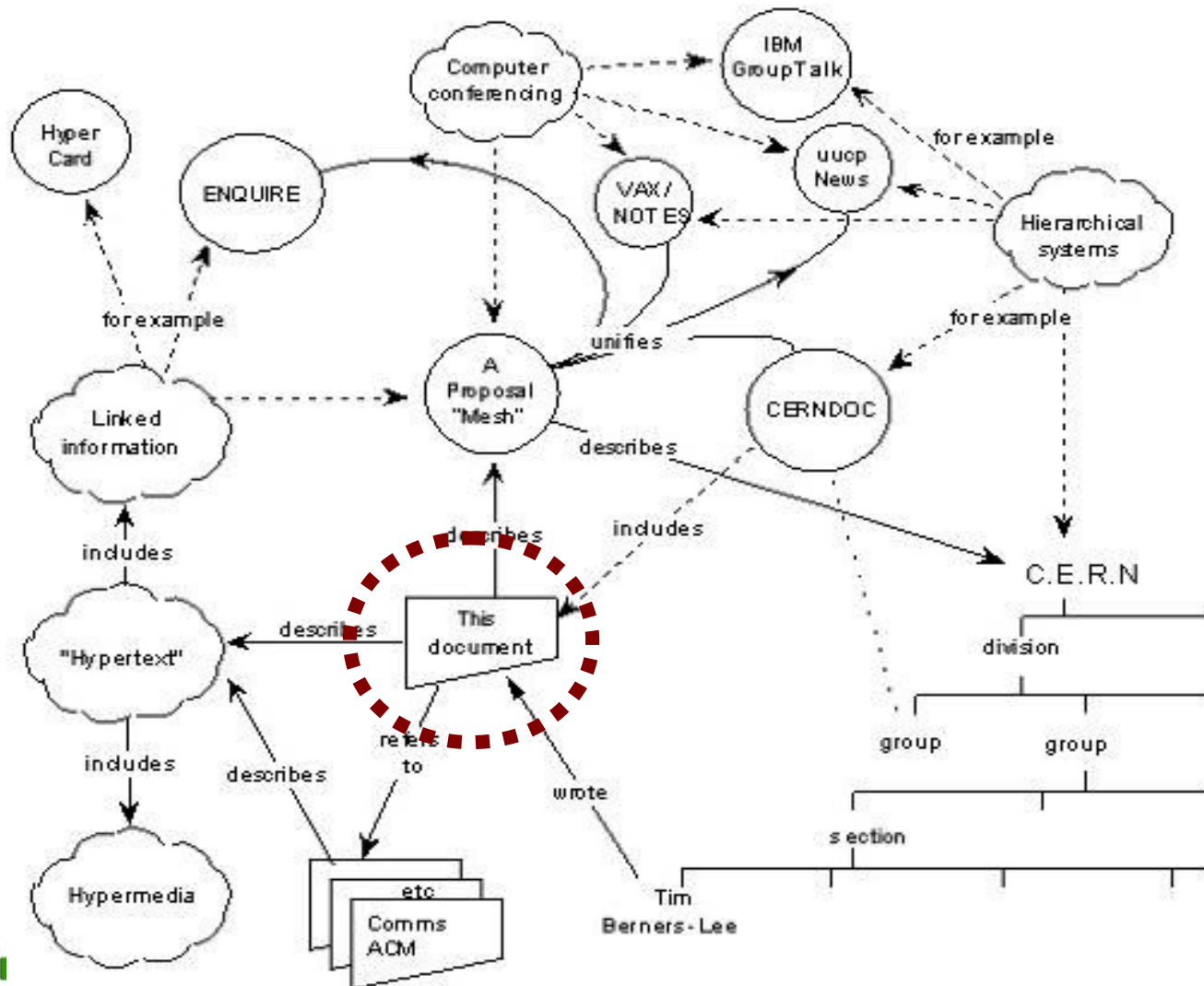
“It's massively successful. It is trivially simple. Massively successful like karaoke - anybody can do it.”
(Ted Nelson 2001)

“Anyone can say anything about anything”
(Tim Berners Lee, 2002)

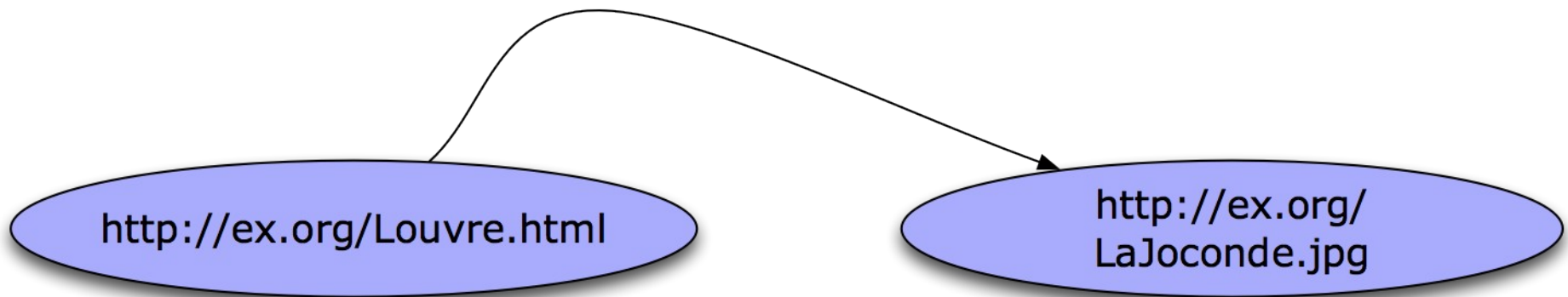
The Document Web

Information Management: A Proposal (TBL, 1989)

- ... twice extended:
- in syntax
- in scope



Resources and Links in the Document Web

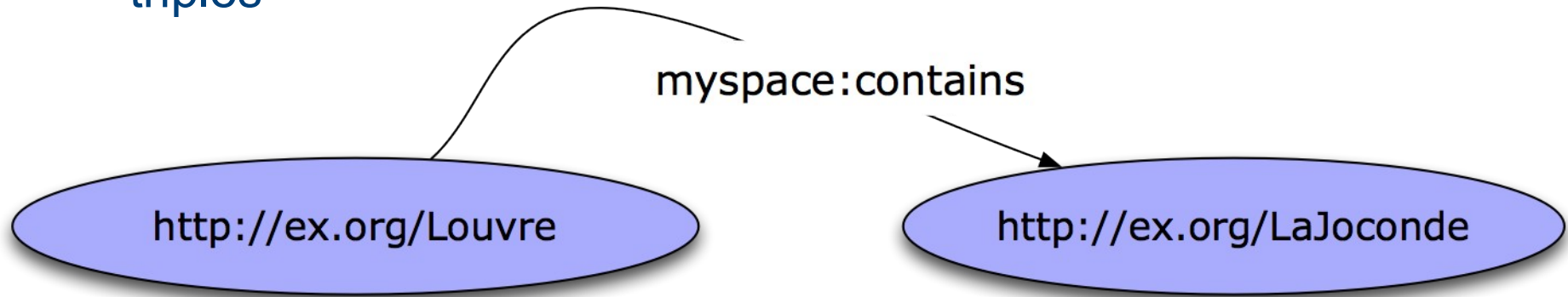


`b`

- We have HTTP URIs to identify resources and links between them – but we are missing a few things!
- What kinds of resources are 'Louvre.html' and 'LaJoconde.jpg'?
 - A machine cannot tell.
 - Humans can: we recognize implied context!
- How exactly do they relate to each other?
 - A machine cannot tell.
 - Humans can: again we recognize implied context!

Syntactically Extending the Document Web (1)

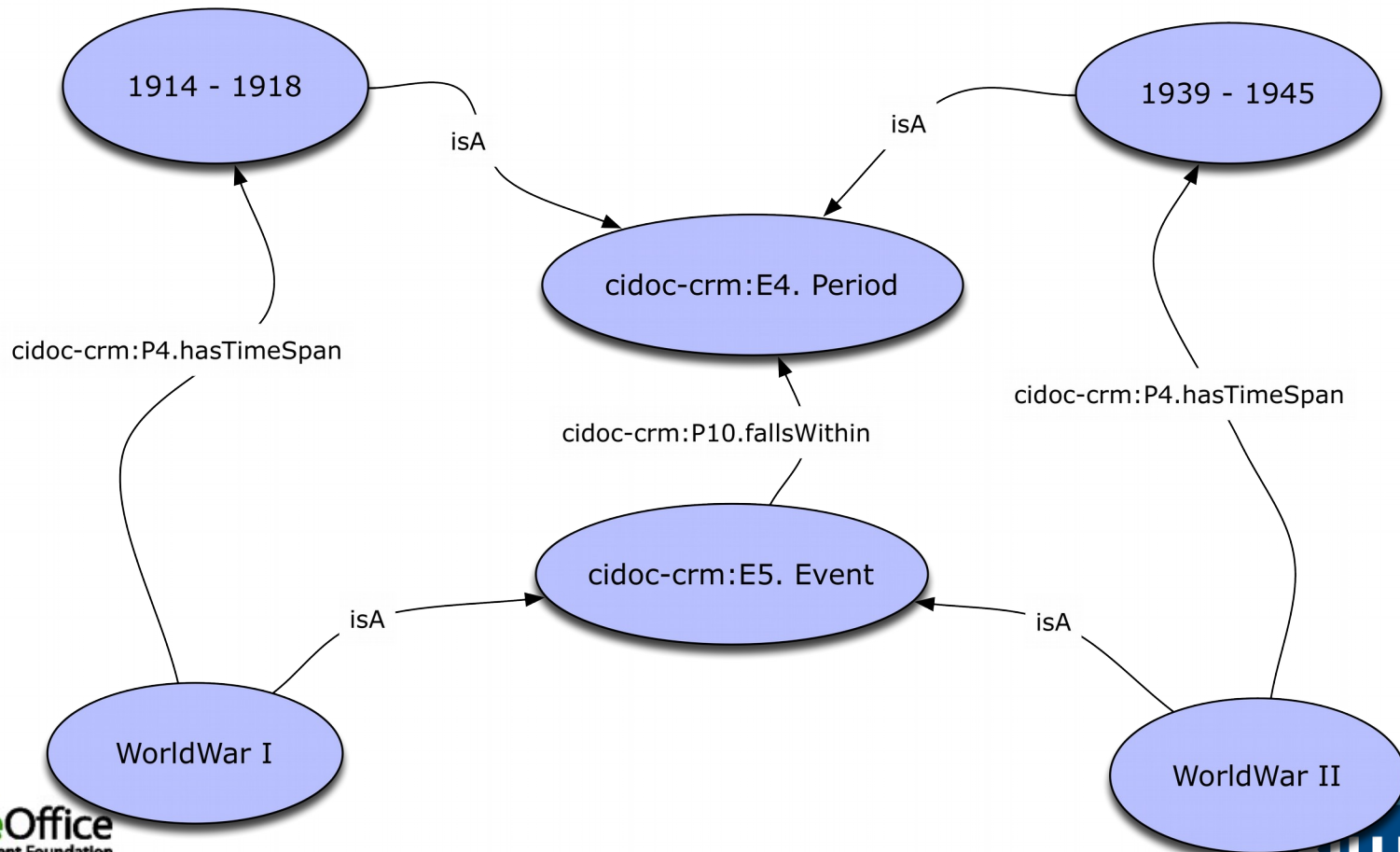
- We add a syntax for making statements on resources: RDF triples



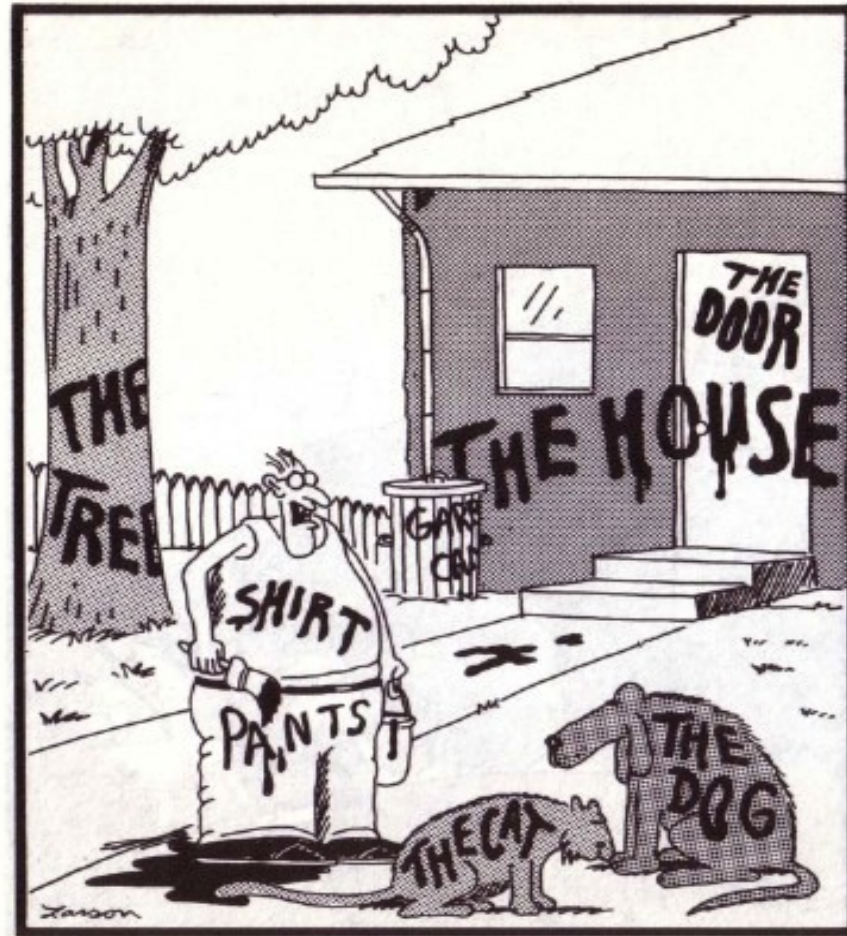
- We add a schema language (RDFS) with elements such as
 - classes ('chair' as instance of chairs),
 - hierarchies of classes and properties (chairs are a subclass of furniture, 'teaches' is a sub-property of 'communicates')
 - inheritance (communication based on language → teaching also is)
 - grammar support for basic inferencing, deterministic logical operations

Syntactically Extending the Document Web: RDF (2)

- And thus are able to establish structures in triple aggregations resulting in lightweight domain ontologies:



Extending the Web in Scope: The Web of Things ... (slightly Mistaken)



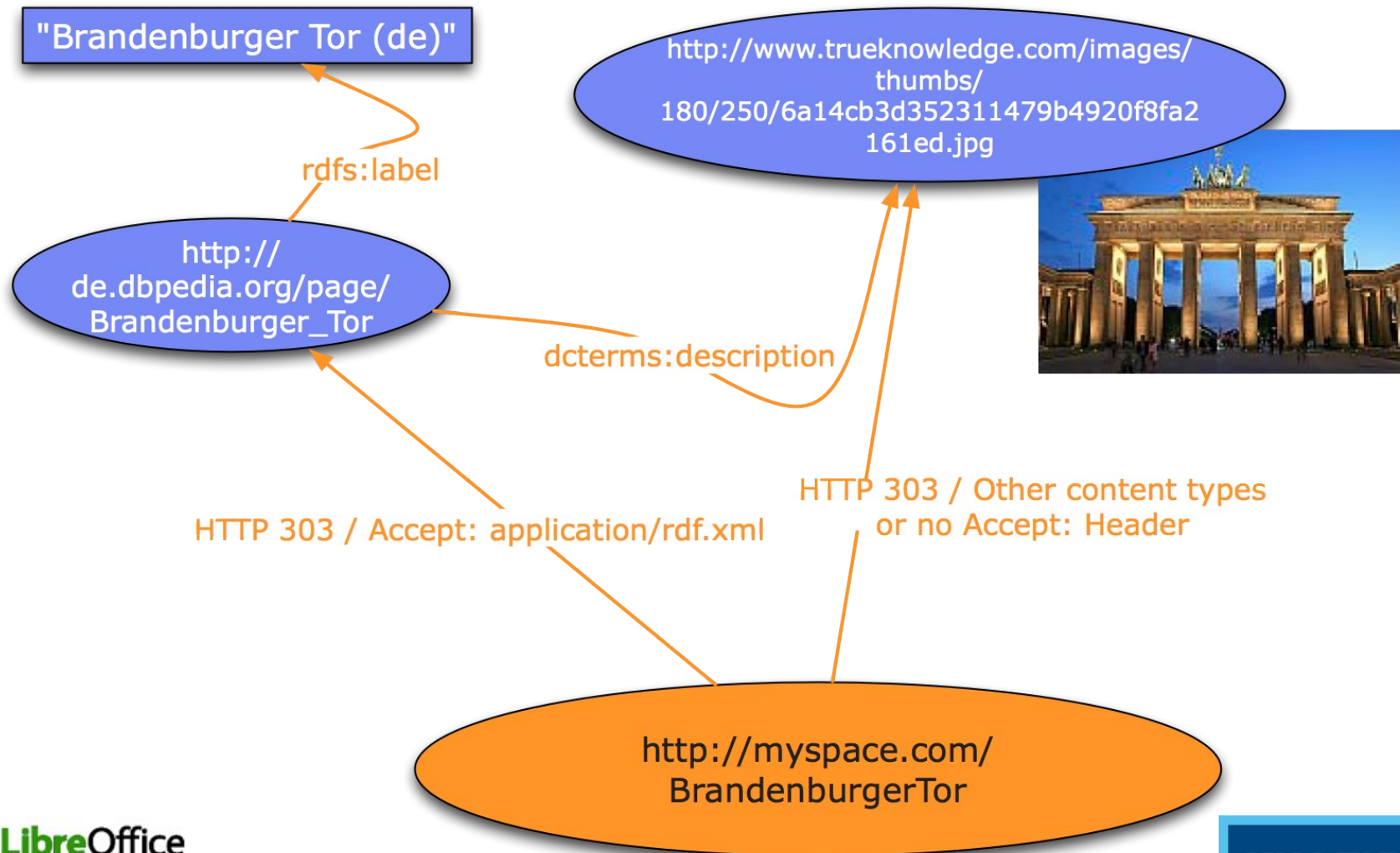
"Now! ... That should clear up
a few things around here!"

- What's wrong with this picture?

Taken from Ronald Carpentier's
Blog at
<http://carpentier.wordpress.com/page/13/>

KU LEUVEN

... and the Way we do extend the Web in scope to make it a 'Web of Things'



And we get ... Linked Data

Linked Data essentials

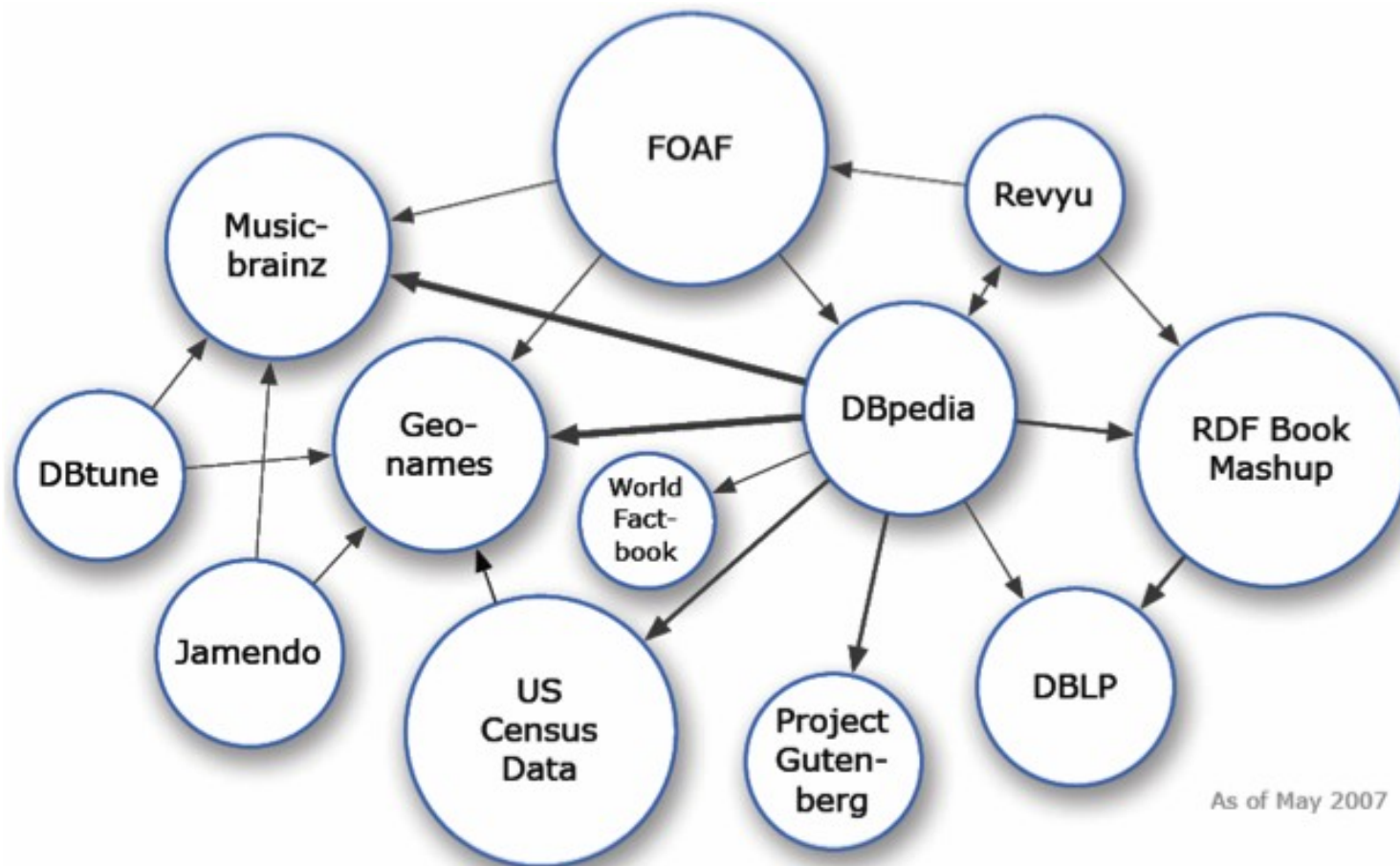
1. Use URIs Standard **Identifiers**
2. Use HTTP URIs Standard **Pointers**
3. Serve useful information using SPARQL, RDF standards
4. Mention URIs of related objects Standards for **Queries** and **Statements**

Link to Context

Copyright © 2008 W3C (MIT, ERCIM, Keio)

[http://www.w3.org/2008/Talks/0617-lod-tbl/#\(4\)](http://www.w3.org/2008/Talks/0617-lod-tbl/#(4))

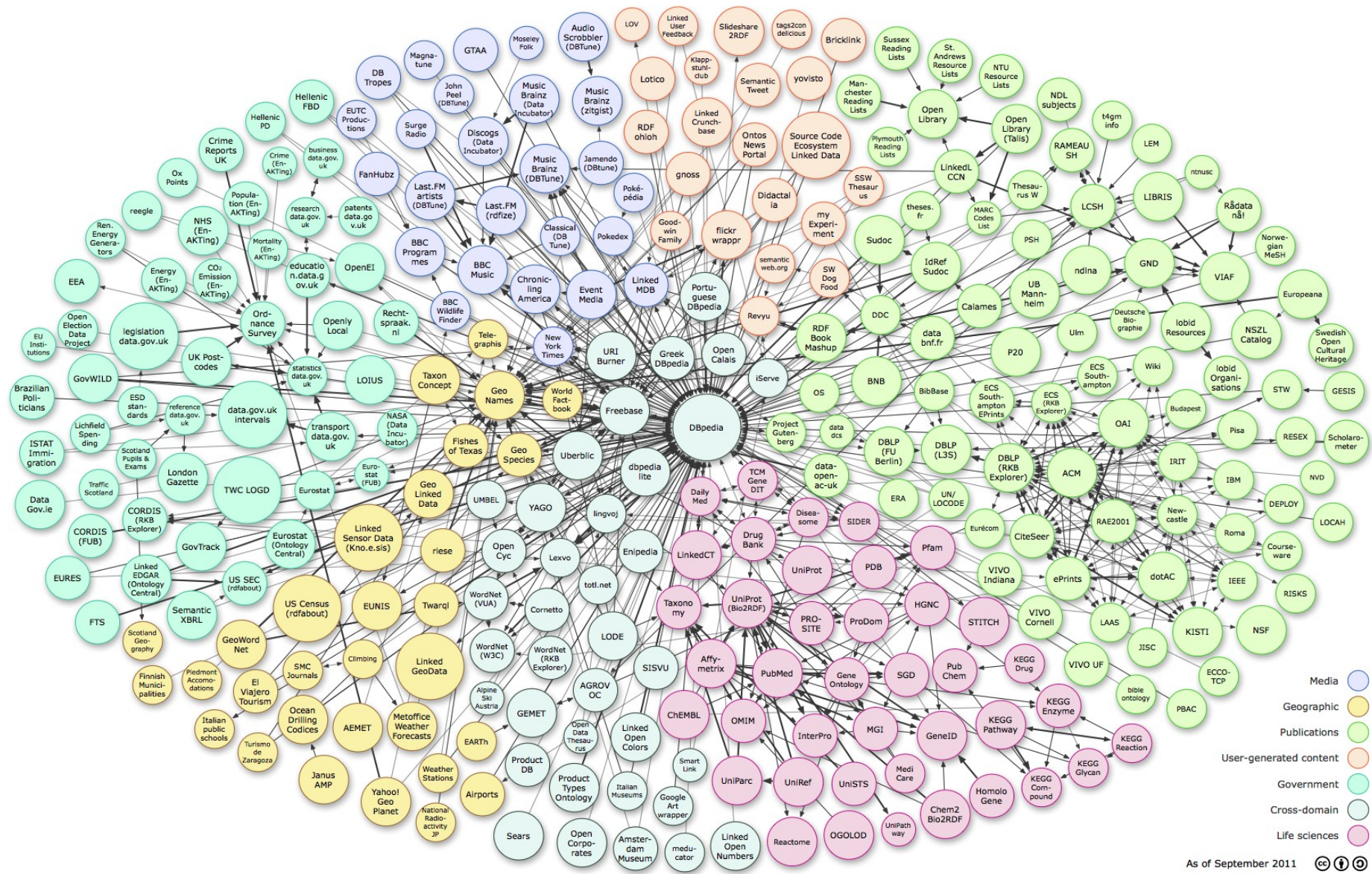
A few Bubbles: 5/2007



- Over 500 million RDF triples

• Around 120.000 RDF links between data sources

And a lot of Bubbles as of last Year



Alternative renderings of LoD ...

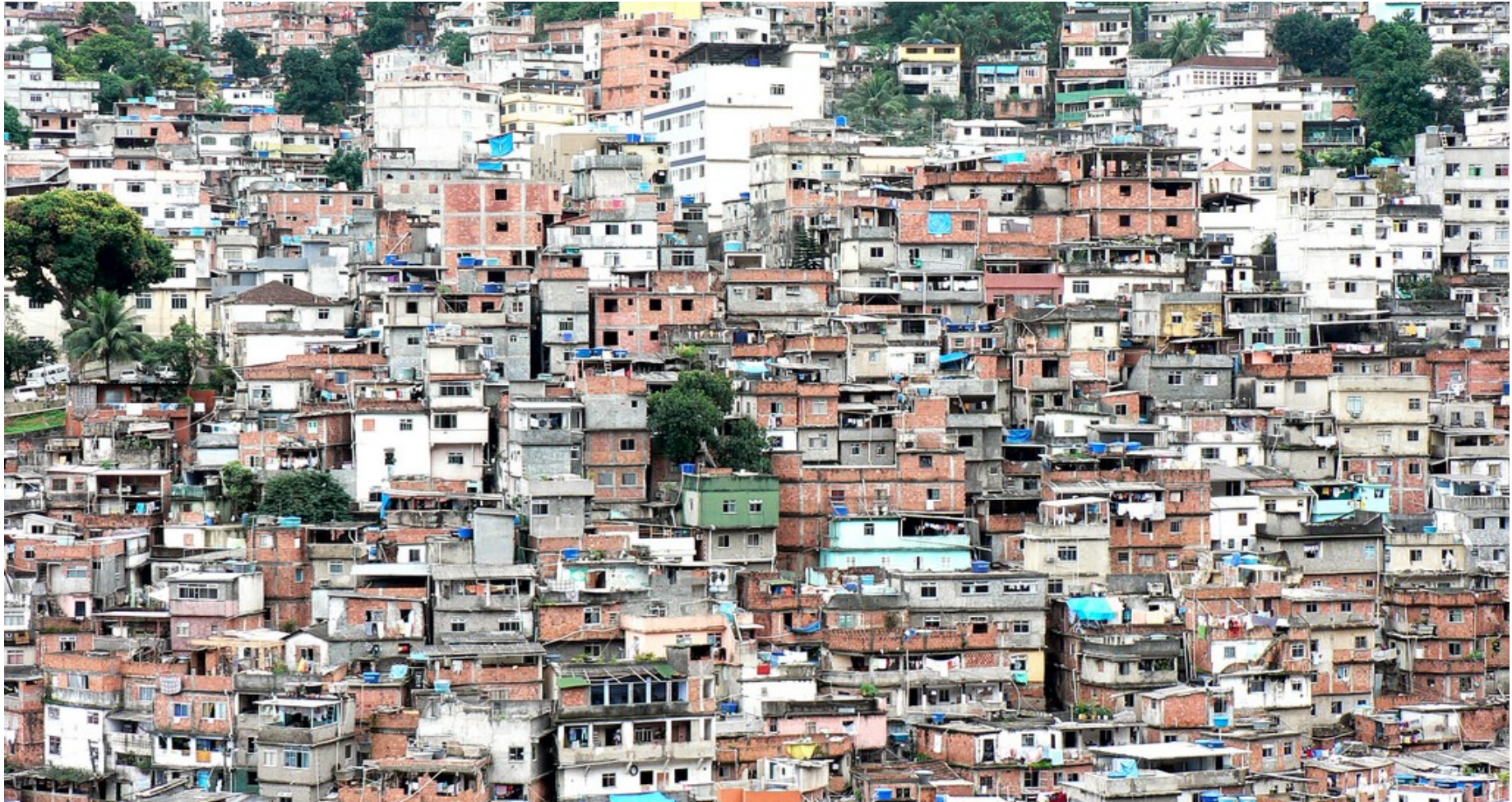
... and an issue!

- Classes
 - <http://inkdroid.org/lod-graph/>
- Instances:
 - <http://zoom.it/Vj6F> or
<http://bgriffen.scripts.mit.edu/www/media/json/thinkers/>
- Vocabularies
 - <http://lov.okfn.org/dataset/lov/>
- Talking about ... persons:
 - <http://ws.nju.edu.cn/falcons/ontologysearch/result.jsp?query=person>

Anyone, Anything ...

- To facilitate operation at Internet scale, RDF is an **open-world framework** that allows **anyone to say anything about anything**. In general, it is not assumed that all information about any topic is available. A consequence of this is that RDF cannot prevent anyone from making **nonsensical** or **inconsistent** assertions, and applications that build upon RDF must find ways to deal with **conflicting sources of information**. (This is where RDF departs from the XML approach to data representation, which is generally quite prescriptive and aims to present an application with information that is well-formed and complete for the application's needs.) (W3C, <http://www.w3.org/TR/2002/WD-rdf-concepts-20020829/>)
- **Issues:** Versioning, provenance, integrity ... and **(partial) semantic redundancy!**

LoD Architecture: the Cloud as Shanty Town



Why not do it this way?



Cathedral Building!



A Finished Cathedral



Schema.org: Order and Discipline

Schema.org: some facts

- Launched June 2, 2011
- By Google, Bing, Yahoo. Later joined by Yandex
- Objective: “create and support a standard set of schemas for structured data markup on web pages”
(<http://blog.schema.org/search?updated-max=2011-12-12T12:10:00-08:00>)
- Controlled core vocabulary set + extensions
- Originally restricted to microdata, RDFa was added in September 2011
- The world as seen at schema.org:
<http://schema.org/docs/full.html>
- Persons in schema.org: <http://schema.org/Person>

Uses of schema.org

- General example source: <http://linter.structured-data.org/examples/>
- Movies
 - <http://www.imdb.com/title/tt0306414/>
 - http://www.rottentomatoes.com/m/the_adventures_of_tintin/

Is there a hidden agenda?

Schema.org – Useful? Innocent? Evil??

Darin Stewart:

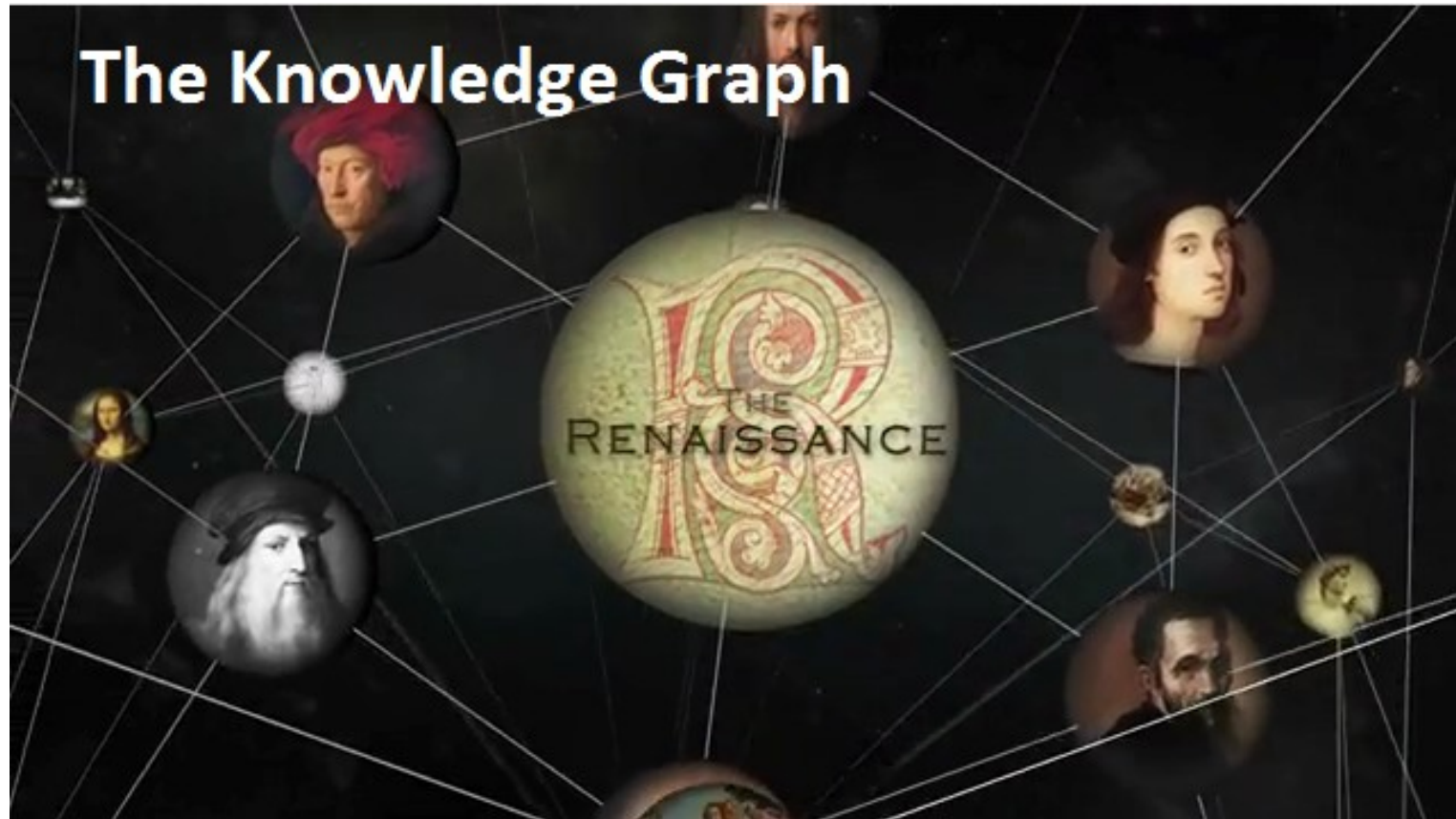
- “Schema.org appears to be Linked Data Lite with extremely limited support for vocabularies outside of the service.”
- “There is a subtle air of intimidation throughout the schema.org announcements and documentation.”
- “Again, I could just be paranoid, but this is Microsoft and Google we’re talking about. Whatever happened to “do no evil?””

(all from <http://blogs.gartner.com/darin-stewart/2011/06/04/schema-org-webmaster-one-stop-or-linked-data-land-grab/>)

What is the motivation?

- Google, Bing, Yahoo and Yandex are fierce competitors.
- What makes them co-operate in a core business area?
- What makes them bend the standards (and be it just slightly)?
- Is schema.org a reaction to Linked open Data??
- Is it thus The Cathedral vs. The Bazaar again (re-read Raymond!) ???

And how does it relate to the Knowledge Graph?



=> <http://manu.sporny.org/2012/google-indexing-schema-rdfa/>

Privatized Censorship?

Privatized Censorship?

- No:
 - Noone commissioned Google to do this.
 - Schema.org just is a radical form of what we may soon be calling ‘vocabulary darwinism’.
- Yes:
 - It is factual censorship, after all.
 - Noone called for this: it isn’t privatized.
 - But the effects are similar in case we just let them go.
 - And they sure fit in Google’s business model.

It may be worse than censorship ...

- Traditional censorship was about preventing people from **publishing** what they had written / created
- The language used / the artistic means as such were **noone's property**
- With schema.org the very means of expression become **oligopolistic property**
- ... schema.org is about **what can be said** (cannot be said) and which statements will be **recognized** (and which ones will simply remain unnoticed)
- Schema.org is about the true currency of the web: **attention!**

Questions?